BY THE NUMBERS

by Duc Lu, AGI

Why Do We Need Higher SNP Density?

The effect of SNP selection on GE-EPD accuracy.

Genomic enhanced expected progeny differences (GE-EPD) are selection tools that are calculated using phenotypic records and DNA information, mostly single nucleotide porlymorphisms (SNP).

The accuracy of GE-EPDs depends on two factors: 1) The proportion of genetic variance explained by the SNPs; 2) The accuracy with which SNP effects are estimated. The latter relies on the heritability of the phenotype and number of animals with phenotypic records and genotypes. This article focuses on the first factor in efforts to detail the effect of SNP selection on GE-EPD accuracy.

In a perfect world, where the training population has a large number of phenotypic records and genotyped animals and the SNP chip contains all causal variants (CVs), the prediction accuracy would be 1 (prediction accuracy ranges from 0 to 1), or 100%. This has been impossible because most traits in livestock production are polygenic, meaning there are many CVs affecting each trait. However, none of the SNPs in the current commercial SNP arrays are CV.

What is included in generic SNP arrays for cattle? The Illumina BovineSNP50 BeadChip has been a popular SNP array used in dairy and beef cattle breeds and was built to have around 54,000 SNPs distributed at relatively even spacing across the bovine genome.

Although none of these SNPs are CV, many are expected to have some level of effect on economically important traits through linkage disequilibrium (LD) between SNP alleles and CVs. Stronger LD means more accurate and robust genomic prediction. Higher SNP density means higher chance for SNPs to be in stronger LD with nearby CVs. The extent of LD in some genomic regions varies depending on breed and breed purity. Thus, the number of SNPs required for a given genomic region may vary among breeds.

SNP in the Illumina BovineSNP50 BeadChip, as well as SNPs in other array designed for multiple breeds, are subject to passing a certain minimum minor allele frequency (MAF) so that a majority of the SNPs are polymorphic, or occur in several different forms including both the homozygous (AA or BB) and heterozygous (AB) states, in all of the target breeds. For that reason, they might have very weak LD with rare or low minor frequency CVs.

Angus Genetics Inc. (AGI) has worked to build and upgrade the available SNP arrays to address the following:

1. Higher SNP density in genomic regions that potentially harbor CVs in the Angus breed. Traits considered include male and female fertility (heifer pregnancy, sperm cell morphology), health or environmental adaptability traits [immune response, pulmonary arterial pressure (PAP), hair shedding], feed efficiency (feed intake, residual feed intake), and meat quality.
2. Low MAF SNP are allowed to enable research of regions that harbor them. Examples of such research include fertility haplotypes and potential SNPs for slickness.

Angus breeders can genotype their animals with either of the two SNP arrays — Angus GS℠ or HD50K™ for Angus — which are then combined to create a SNP set with more than 85,000 SNPs on autosomes, or any chromosome that is not a sex chromosome (X or Y).

All genotyped animals in the Angus database go through a process called genotype imputation to have 85,000 SNP density regardless of the SNP array. All 85,000 SNPs enter a series of analyses aiming for selecting a subset of SNPs that could potentially provide more information than the current SNP set used in Angus's weekly genetic evaluation. While the current SNP set is very good, a more targeted SNP set could boost

prediction accuracy for different traits, especially those that are new and lowly recorded, e.g. PAP and hair shedding.

In that process, AGI has tested several subsets of the 85,000 SNPs and have observed preliminary outcomes including prediction accuracy increases between 5% and 10% for PAP and hair shedding.

Subsetting SNPs from DNA sequence could yield further improvement, because the sequence contains millions of SNPs and potential CVs. The AGI team has been sequencing influential animals and phenotype-specific animals to serve this project aiming for a new set of SNPs containing putative CVs and/or SNPs that are in strong LD with CVs to make genomic prediction for the Angus breed more accurate and more robust.

Sequencing the large number of necessary animals is costly, thus AGI is stretching the project over several years to allocate financial resources and continue to target the next generation of highly used sires.

The AGI team will continue to drive this research forward. Angus members should keep an eye out for potential opportunities to participate in this research in the future. AJ

Glossary

DNA sequence is the order of nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T) in an individual's DNA. These nucleotides exist in pairs. There are approximately three billion of such pairs in the bovine genome, but only 1% of them are believed to be polymorphic. Each of those pairs can be referred to as a DNA variant or ***single nucleotide polymorphism (SNP)***.

A ***causal variant (CV)*** is a DNA variant that causes differences in phenotype.

A ***SNP array*** is a set of SNPs with which animals are genotyped, e.g., Angus GS, HD50K for Angus.

Linkage disequilibrium (LD) is a phenomenon where alleles occur together more often than can be accounted for by chance because of their physical proximity on a chromosome.

There are two alleles at each SNP, one of which normally exists at a higher frequency than the other in the population. ***Minor allele frequency (MAF)*** refers to the frequency of the less frequent allele, which often differs among breeds.

Imputation is a statistical process of inferring missing SNP calls based on SNPs existing in a reference population.